

本周周报

解聪

2013.11.25-2013.12.1

本周工作

一、各维度互信息的计算

本周继续对企业数据进行分析。企业数据中除去基本信息(如企业的名称,地理位置等),还有大量的比较专业的维度,基本是数值型的,一共 77 个。

首先计算维度各自分部的熵。 $H = -\sum P \ln P$ 。

得到计算结果中,

熵比较大的属性为: V209 工业销售产值, V210 全部从业人员年平均人数, V211 工业增加值, F325 主营业务收入, F349 本年应付工资总额, F350 主营业务应付工资总额。

这些维度属于比较好理解的维度,同时各企业的差异比较大。

熵比较小的维度有: V212 其中新产品产值, F1 短期投资, F302 国家资本, F392 集体资本, F7 投资收益, F343 补贴收入, F3621 投资活动产生的现金流入。

这些维度比较专业,不是所有企业都有值大部分企业是 0,因此分布较稀疏,从而熵较小。

再对两两维度计算其相对熵又称为 Kullback-Leibler 距离: $d_{KL} = \sum P \ln(P/Q)$ 。

由于相对熵是不对称的,所以后来选择计算其 Jensen-Shannon 距离:

$d_{JS} = 0.5 * (\sum P \ln(2P/(P+Q)) + \sum Q \ln(2Q/(P+Q)))$

计算得到 77*77 的矩阵:

	V207	V212	V209	V213	V210	V211	F304	F1	F390	F301
V207	0	380.5697	0.21511	186.6959	23.07637	4.171196	5.466116	417.3926	21.39991	26.
V212	380.5697	0	384.1094	128.66	518.7514	402.7267	348.883	20.96639	306.1707	29
V209	0.21511	384.1094	0	189.4587	22.31036	3.987465	5.755796	420.9549	21.75809	27.
V213	186.6959	128.66	189.4587	0	301.2261	204.0075	159.7942	155.5667	130.1533	11
V210	23.07637	518.7514	22.31036	301.2261	0	16.4699	40.56941	558.2799	72.46959	83.
V211	4.171196	402.7267	3.987465	204.0075	16.4699	0	11.43868	439.9564	27.64774	32.
F304	5.466116	348.883	5.755796	159.7942	40.56941	11.43868	0	384.8041	12.67126	17.
F1	417.3926	20.96639	420.9549	155.5667	558.2799	439.9564	384.8041	0	340.553	328
F390	21.39991	306.1707	21.75809	130.1533	72.46959	27.64774	12.67126	340.553	0	3.8
F305	26.84172	294.705	27.22847	119.703	83.82339	32.84557	17.39707	328.7594	3.805415	
F306	42.87489	266.4961	43.99162	98.65988	115.2408	53.53017	31.32853	299.3589	16.05722	11.
F308	5.813631	348.1431	6.082978	159.508	40.80413	11.8991	0.507045	384.0456	12.26148	17.
F391	361.812	10.08525	365.2841	115.3939	498.5428	383.7496	330.5682	31.10614	288.7572	277
F309	40.46409	253.5087	41.67772	89.97353	114.9923	51.02579	27.77746	286.2016	20.41507	16.
F310	34.02245	262.5477	34.78293	95.52687	104.614	43.46486	23.00073	295.6075	16.25369	12.
F311	51.41743	238.8661	52.70564	78.10972	128.7804	62.51413	36.9516	271.0493	21.1254	16.
F312	82.20342	206.8478	83.97175	57.77166	170.7295	94.26129	61.25634	237.6738	40.98082	35.
F313	62.72435	224.4595	64.23429	68.96475	144.4107	74.339	44.49449	256.1631	28.18646	23
F314	39.42689	255.223	40.62582	91.1871	113.5049	49.95104	27.1061	287.9285	19.83276	16.
F315	122.6122	122.6122	122.6122	122.6122	122.6122	122.6122	122.6122	122.6122	122.6122	122

可以发现有的维度的距离很小,如 V209 工业销售产值和 V207 工业总产值,通过调查发现这两个维度基本是一样的,说明这些数据存在比较明显的冗余信息。

76 个数值维度具体为:

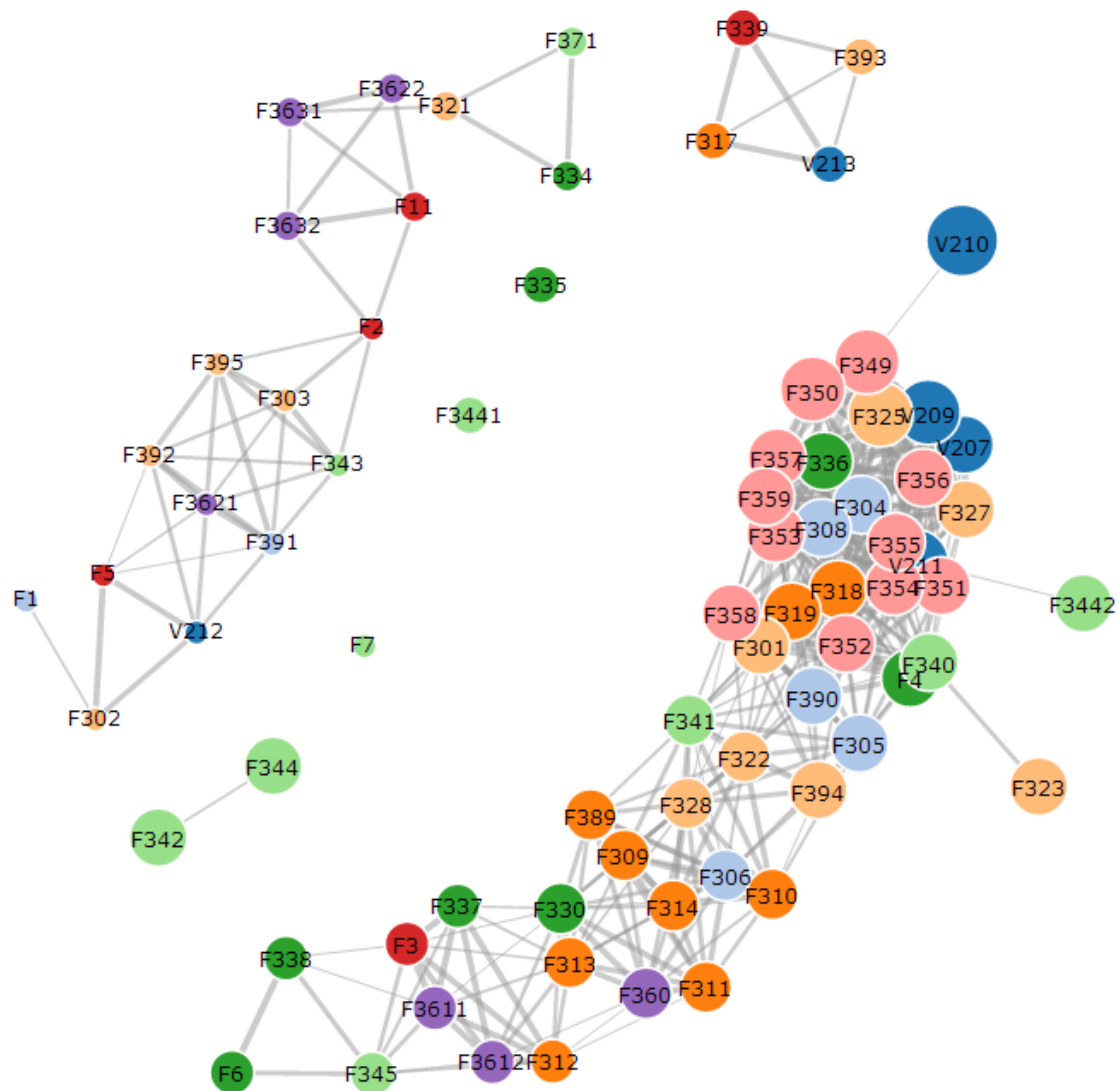
v207 工业总产值	f389 其中：应付帐款	f342 营业利润	f354 本年进项税额
v212 其中新产品产值	f321 长期负债合计	f7 投资收益	f355 本年销项税额
v209 工业销售产值	f322 负债合计	f343 补贴收入	f356 工业中间投入合计
v213 其中出口交货值	f323 所有者权益合计	f371 营业外收入	f357 其中：直接材料
v210 全部从业人员年平均人数	f301 其中：实收资本	f344 利润总额	f358 制造费用中的中间投入
v211 工业增加值	f302 1. 国家资本	f3441 亏损总额	f359 管理费用中的中间投入
f304 流动资产合计	f392 2. 集体资本	f3442 利税总额	f360 营业费用中的中间投入
f1 其中：短期投资	f393 3. 法人资本	f345 应交所得税	f3611 经营活动产生的现金流入
f390 应收帐款	f394 4. 个人资本	f2 广告费	f3612 经营活动产生的现金流出
f305 存货	f395 5. 港澳台资本	f5 研究开发费	f3621 投资活动产生的现金流入
f306 其中：产成品	f303 6. 外商资本	f339 劳动、失业保险费	f3622 投资活动产生的现金流出
f308 流动资产年平均余额	f325 主营业务收入	f3 养老保险和医疗保险费	f3631 筹资活动产生的现金流入
f391 长期投资	f327 主营业务成本	f11 住房公积金和住房补贴	f3632 筹资活动产生的现金流出
f309 固定资产合计	f330 主营业务税金及附加	f349 本年应付工资总额	
f310 固定资产原价	f334 其他业务收入	f350 主营业务应付工资总额	
f311 其中生产经营用	f335 其他业务利润	f351 本年应付福利费总额	
f312 累计折旧	f328 营业费用	f352 主营业务应付福利费总额	
f313 其中：本年折旧	f336 管理费用	f353 本年应交增值税	
f314 固定资产净值年平均余额	f337 其中：税金		
f317 无形资产	f338 财产保险费		
f318 资产总计	f4 办公费		
f319 流动负债合计	f6 职工教育费		
	f340 财务费用		
	f341 其中：利息支出		

对于生成的距离矩阵，简单起见，设定阈值去除了矩阵中距离比较大的元素，只研究互信息较大的元素。

使用力引导布局可视化简化后的矩阵，每个维度作为图中的一个节点，使用 D3 实现。具体可以参考上面的维度表。

每个节点的大小表示该维度单独分布的熵大小。

可视化结果见下图：



其中每个节点的颜色按照维度编号来编码（比如 F334 和 F335 定义为一类）。具体为：“F30” 开头，“ F31” 开头， “ F32” 开头， “F33” 开头， “ F34” 开头， “ F35” 开头， “F36” 开头， “ V2” 开头， 其他

图中表现了各维度本身，以及各维度之间是具有较明显的模式的。

1. 可以看到图中的粉色的点是聚到一起的，而且其半径较大，说明 F35 系列的维度互信息较大，而且不确定性较大。比如：主营业务应付工资总额，本年应付福利费总额，工业中间投入合计等。
2. 左上方的连通图的节点普遍比较小，说明个维度单独分布熵比较小。查看数据库后发现这些维度都是比较稀疏的维度。大部分企业都是默认 0 值，所以其分布大致相似。但其互信息较大并不是特别能代表两个维度之间关系很密切。
3. 右上四个点分别是：V213 出口货值, F393 法人资本, F339 劳动、失业保险费, F317 无形资产。感觉是毫不相关的四个维度，至于为什么聚在一起，还有待调查。

利用这个视图，可以帮助用户进行维度的选取。避免选择混乱程度较大的维度对人群进行划分。

同时，在已选定一个维度的基础上，可以利用该视图，避免选取互与现有维度信息较大

的维度。

二、对单个数值维度的划分

上周使用改进的迭代算法，效果感觉不是很好。本周还在继续查阅相关文献中。

三、专利

周五找陈律师，医学和音符的专利还是又很多地方要改

下周工作：

继续探索企业数据维度之间的关联性。并将其与马赛克图相结合。

继续修改专利。